Generación automática de resúmenes personalizados

Ignacio Acero, Matías Alcojor, Alberto Díaz, José María Gómez

Departamento de Inteligencia Artificial, Escuela Superior de Informática, Universidad Europea-CEES,
28670, Villaviciosa de Odón, Madrid
Tel.: (34) 91 6167142 ext. 671. Fax: (34) 91 6168265
nachoa@telcom.es, alcojor@retemail.es
{alberto, jmgomez}@dinar.esi.uem.es

Manuel Maña

Departamento de Informática, Universidad de Vigo,
Campus As Lagoas s/n, 32004, Orense, Spain.
Tel.: (34) 988 387014. Fax: (34) 988 387001
mjlopez@uvigo.es

Resumen. En la actualidad los servicios de información presentes en la Web y en particular los periódicos digitales ofrecen a los usuarios una selección de documentos basada en criterios bastante simples que lleva a los usuarios a recibir una gran cantidad de información irrelevante. Nuestro trabajo pretende disminuir la sobrecarga de los usuarios de dos maneras: aportando un modelo de usuario más completo que permita definir mejor los intereses de los usuarios y construyendo un generador de resúmenes automático que permita al usuario detectar las noticias que realmente le interesan mediante la visualización de resúmenes adaptados a su modelo. Se han obtenido resultados alentadores en una evaluación de los distintos tipos de resúmenes para varios modelos de usuario. El trabajo está enmarcado dentro del sistema Hermes, un enviador personalizado de noticias que maneja información en inglés y en español.

1. Introducción

Los servicios personalizados de información son cada vez más populares en la Web. Los periódicos digitales más importantes ofrecen a sus usuarios la posibilidad de recibir por correo electrónico una selección de las noticias del día. La mayoría de los periódicos envían todas las noticias (o sólo los titulares) de una o varias secciones que el usuario ha seleccionado previamente, o también aquellas que contienen ciertas palabras elegidas por el usuario.

Una definición tan simple de los intereses de un usuario puede hacer que mucha de la información recibida sea realmente irrelevante. La creación de modelos de usuario que mejoren la definición de esos intereses junto con la integración de tareas de clasificación de texto permite mejorar la selección de las noticias relevantes para cada usuario. Ejemplos representativos de sistemas de acceso a la información que integran este tipo de técnicas son WebMate [5], News Dude [3] y SIFT [18].

En todo caso, aunque la selección se mejore, el usuario recibe un conjunto de noticias en su correo todos los días de las cuales es probable que no todas le interesen, o unas le interesen más que otras, o incluso le interese conocer algunas en profundidad y otras superficialmente. Para evitar que el usuario tenga que leer todas las noticias, una mejora que puede introducirse, es incluir un resumen de cada noticia, en vez de enviar la noticia completa. De esta manera, el usuario lee solamente el resumen de las noticias y en caso de que le interese, puede acceder a la noticia completa.

Reemplazar la noticia completa por un resumen es una mejora significativa, pero podríamos mejorar aún más estos sistemas si el resumen fuera personalizado para cada usuario. Así cada usuario podría leer un resumen adaptado a sus preferencias definidas en su modelo de usuario.

Este trabajo está incluido en el proyecto Hermes[1], el cual tiene como objetivo la personalización inteligente en servicios de envío de noticias mediante la integración de técnicas de análisis automático del contenido textual y modelado de usuario con capacidades bilingües. El sistema de generación de resúmenes es un módulo dentro del proyecto. En este trabajo describimos el funcionamiento para el idioma español, pero se ha desarrollado de tal manera que la mayor parte del mismo es independiente del idioma y el coste de adaptación al inglés es, por tanto, mínimo.

2. Generación de resúmenes

En la actualidad, buena parte de la información que leemos a diario nos llega a través de medios electrónicos. El amplio uso que de estos medios se está realizando para la difusión de información conlleva una excesiva carga para el lector, que hace cada vez más necesaria la disponibilidad de herramientas que permitan resumir estos textos.

Por generación automática de resúmenes de texto entendemos el proceso por el cual se identifica la información sustancial proveniente de una fuente (o varias) para producir una versión abreviada destinada a un usuario particular (o grupo de usuarios) y una tarea (o tareas) [11]. Bajo esta definición se agrupan diferentes tipos de resúmenes que pueden clasificarse atendiendo a su propósito, enfoque y alcance [8].

Atendiendo al *alcance*, el resumen puede limitarse a un único documento o a un conjunto de ellos que traten sobre el mismo tema.

Según su *propósito*, esto es, atendiendo al uso o tarea al que están destinados, los resúmenes se clasifican en:

- *Indicativos*, si el objetivo es anticipar al lector el contenido del texto y ayudarle a decidir sobre la relevancia del documento original,
- *Informativos*, si pretenden sustituir al texto completo incorporando toda la información nueva o trascendente, y
- *Críticos*, si incorporan opiniones o comentarios que no aparecen en el texto original.

Finalmente, atendiendo al *enfoque*, podemos distinguir entre resúmenes:

- *Genéricos*, si recogen los temas principales del documento y van destinados a un grupo amplio de personas, y
- *Adaptados al usuario*, si el resumen se confecciona de acuerdo a los intereses (i.e. conocimientos previos, ámbitos de interés o necesidades de información) del lector o grupo de lectores al que va dirigido.

Es claro que, sobre todo para niveles de compresión altos, un resumen que no tenga en cuenta las necesidades del usuario puede ser demasiado general como para ser útil. En concreto, en un entorno de recuperación de información, ya ha sido demostrada la superioridad de los resúmenes adaptados a la consulta realizada por el usuario [12]. En otros ámbitos relacionados, como el filtrado o los servicios personalizados de información, en los que el sistema puede disponer de un mayor conocimiento sobre las preferencias de los usuarios, cabría esperar una selección más adecuada de los contenidos.

Ante la variedad de tipos y dominios de los documentos disponibles, las técnicas de selección y extracción de frases resultan muy atractivas por su independencia del dominio y del idioma. El método se centra en una fase de análisis en el que se identifican los segmentos de texto (frases o párrafos, normalmente) que contienen la información más significativa. Durante esta fase se aplica un conjunto de heurísticas a cada una de las unidades de extracción. El grado de significación de cada una de ellas puede obtenerse mediante combinación lineal de los pesos resultantes de la aplicación de dichas heurísticas. Éstas pueden ser *posicionales*, si tienen en cuenta la posición que ocupa cada segmento dentro del documento, *lingüísticas*, si buscan ciertos patrones de expresiones indicativas, o *estadísticas*, si incluyen frecuencias de aparición de ciertas palabras.

El resumen resulta de concatenar dichos segmentos de texto en el orden en que aparecen en el documento original [10]. Los posibles problemas de inconsistencia en el resumen resultante constituyen el principal inconveniente de esta aproximación. Una forma de paliar este problema es utilizar el párrafo en lugar de la frase como unidad de extracción [14], esperando que al proporcionar éste un

[1]Este proyecto ha sido financiado parcialmente por el Ministerio de Ciencia y Tecnología (PROFIT, 2000/020)

contexto más amplio se mejoren los problemas de legibilidad.

Otro conjunto de técnicas, que podemos denominar *ricas en conocimiento*, se caracteriza por utilizar métodos de comprensión profunda del texto [9]. Esta aproximación puede conducir a crear sistemas que evitan algunos problemas, como los mencionados de inconsistencia. Sin embargo, estos sistemas necesitan gran cantidad de conocimiento sobre el dominio. Además, estos dominios son necesariamente muy restringidos y de características conocidas. Como consecuencia, los sistemas son poco flexibles y de difícil adaptación a otros casos de aplicación u otros idiomas.

3. *Modelado de usuario*

El modelo sirve, como se indica en la introducción, para filtrar las noticias que más le interesan a un usuario [6].

El modelo de usuario está diseñado para representar los intereses del usuario desde varios puntos de vista [1]. Este perfil describe las necesidades de información del usuario, es decir, lo que el usuario realmente está buscando.

El modelo de usuario almacena 3 tipos de información:

- Información general (nombre, login, password, dirección de correo electrónico).
- Información sobre sus preferencias (días de la semana que desea recibir mensajes, máximo número de noticias por mensaje, desactivación temporal del servicio).
- Información sobre los intereses del usuario: (1) secciones, (2) categorías generales, (3) términos.

La información sobre las preferencias es muy útil para los usuarios. La posibilidad de fijar un máximo número de noticias por mensaje y el hecho de poder desactivar temporalmente el servicio evitan la sobrecarga de información. Establecer un número mínimo de noticias ha sido considerado contraproducente porque puede llevar a la inclusión de información no relevante si no se alcanza dicho número con la información relevante.

Como se ha descrito en la introducción, los servicios de noticias normalmente realizan la personalización de sus envíos utilizando información sobre las secciones y palabras claves seleccionadas por el usuario. Nuestro modelo soporta ambos aspectos.

Por una parte, los usuarios pueden seleccionar sus secciones preferidas. Las secciones de un periódico son un conjunto definido de categorías basadas en el contenido de las noticias. Este conjunto de 8 categorías procede de la organización tradicional de un periódico. Los usuarios pueden asignar un peso a cada sección, para dar mayor importancia a las noticias que pertenezcan a dichas secciones. Ejemplos de secciones en un periódico son “Internacional” o “Cultura”.

Por la otra, los usuarios también pueden introducir un conjunto de términos y asignar un peso a cada uno de ellos que represente su importancia para los intereses del usuario.

Nuestro modelo permite otro sistema de referencia adicional utilizando un sistema diferente de categorías. Los usuarios de Internet están familiarizados con los sistemas de categorías utilizados en los motores de búsqueda, por ello se ha elegido el sistema de categorías de Yahoo! España como sistema alternativo para representar los intereses de un usuario. Al ser un sistema de mayor cobertura que las secciones de un periódico constituye una buena segunda opinión. Se ha elegido el primer nivel de 14 categorías para representar esta información. Los usuarios pueden asignar un peso a cada categoría, para indicar su interés en ella.

La existencia de muchos métodos de selección puede llegar a confundir al usuario, haciéndole tener una idea borrosa del funcionamiento del sistema. Por ello, existe un nivel adicional de especificación de los intereses del usuario. Cada uno de los 3 sistemas de referencia del modelo (secciones, categorías y términos) tienen un peso que representa la importancia para el usuario de cada uno de ellos. Por ejemplo, si el peso de las secciones es bajo y el peso de las categorías es alto, los valores de relevancia procedentes de la categorización automática de las noticias con respecto a las categorías tendrá mayor importancia para la selección de las noticias. De esta manera, cada una de las 3 dimensiones del perfil de usuario puede ser definida y controlada por el usuario, constituyendo un mecanismo fino de ajuste para obtener una caracterización flexible de sus intereses.

El modelo puede ser editado por el usuario: se pueden modificar los pesos de secciones y términos, y se puede borrar o añadir términos.

En el proceso de selección se aplica categorización de texto [4, 16] con las categorías respecto de las noticias y recuperación de información [2] con los términos con respecto a las noticias. Adicionalmente las noticias son procesadas para comprobar si pertenecen a alguna de las secciones seleccionadas por el usuario. Los diferentes resultados obtenidos se integran usando el nivel de interés asignado por el usuario a los diferentes sistemas de referencia.

4. Generación de resúmenes personalizados.

Para la confección de los resúmenes nuestro sistema utiliza tres heurísticas de selección de frases. Las dos primeras se utilizan para la obtención de resúmenes generales mientras que la tercera se refiere a la personalización de los mismos. Nuestra investigación se centra principalmente en las técnicas de personalización.

Para generar los resúmenes utilizaremos la combinación ponderada de las distintas heurísticas. Vamos a describir cada una de ellas por separado y después comentaremos como modificamos los diferentes parámetros para obtener resúmenes generales o personalizados.

Las tres heurísticas tienen un objetivo común y es dar puntuaciones a cada una de las frases del texto objeto del resumen, para más tarde poder elegir las más relevantes.

4.1 Heurística de posición

Esta heurística consiste básicamente en dar mayor puntuación a las cinco primeras frases de un texto [7].

En dominios periodísticos, el título y las primeras frases de un texto dan una idea aproximada al lector del contenido del texto que va a leer a continuación. Por esta razón asignaremos los siguientes pesos a las N primeras frases del documento que estemos resumiendo. Un valor típico de N podría ser 5 y los pesos asignados podrían ser:

Peso frase 1: 1.0
Peso frase 2: 0.99
Peso frase 3: 0.98
Peso frase 4: 0.95
Peso frase 5: 0.90

4.2 Heurística de palabras clave

Cada texto tiene un número de palabras clave, que son bastante representativas de su contenido. Esta heurística consiste en extraer las M palabras más significativas de cada texto y comprobar a continuación, cuantas de esas palabras clave se encuentran en cada frase. De esta forma asignaremos mayor peso a las frases que contengan mayor número de palabras clave del texto [17].

Para obtener las M palabras más relevantes de cada noticia indexamos todas las noticias obteniendo así el peso de cada palabra en cada documento utilizando el método tf · idf [15]. Seleccionamos así las ocho palabras que más peso tengan para cada documento.

Para obtener el peso de la frase en el documento se divide el número de palabras clave del documento que aparecen en la frase por el número total de palabras de la frase:

4.3 Heurística de personalización

El objetivo de esta heurística consiste en potenciar aquellas frases que tengan mayor relevancia para un modelo de usuario dado, con el fin de personalizar el resumen. En lugar de obtener una idea general del texto resumido, orientamos la elección de frases de tal forma que se elijan aquellas que tengan mayor similitud con las preferencias del usuario.

El potencial de personalizar un resumen es alto, ya que un documento que resumido (de forma general) podría ser descartado, no lo será si se acierta a elegir las frases adecuadas para despertar el interés de dicho usuario.

El cálculo de los pesos para las frases se realiza de la siguiente manera. Del modelo de usuario obtenemos la información con respecto a los pesos que el usuario ha asignado a sus categorías y a sus términos personales. También extraemos del modelo los términos que representan cada categoría así como los términos que el usuario haya definido. Con toda esta información calculamos la similitud existente entre el modelo y la frase, asignando un peso a la frase de acuerdo con la siguiente similitud:

$$PesoFrase = \frac{pCat \cdot sim(frase, Categorías) + pTer \cdot sim(frase, Térms)}{pCat + pTerm}$$

Donde:

$$sim(frase, Categorías) = \frac{\sum_{i=1}^{i=n} sim(frase, tc_i) \cdot pc_i}{\sum_{i=1}^{i=n} pc_i}$$

$$sim(frase, Términos) = \frac{\sum_{i=1}^{i=n} sim(frase, t_i) \cdot pt_i}{\sum_{i=1}^{i=n} pt_i}$$ Siendo, *pCat* el peso general de las categorías, *pTerms* el peso general de los términos, tc_i los términos que identifican a la categoría i, pc_i el peso asignado a la categoría i, t_i el término i y pt_i el peso asignado al término i.

4.4 Combinación de las tres Heurísticas.

Para poder combinar las tres heurísticas y obtener así un solo peso para cada frase utilizaremos la siguiente ecuación:

$$PesoTotalFrase = \frac{(\alpha \cdot pesoH1) + (\beta \cdot pesoH2) + (\gamma \cdot pesoH3)}{\alpha + \beta + \gamma}$$

Los parámetros α, β y γ nos sirven para dar más importancia a una heurística que a otra.

Los cálculos de similitud que realiza nuestro sistema se basan en el modelo del espacio vectorial [15], utilizando para la representación de textos, vectores de pesos de términos. Para su obtención, se eliminan las palabras más frecuentes usando una lista de parada estándar y las restantes se reducen a una forma canónica usando el extractor de raíces de Porter adaptado al español [13].

5. Evaluación

En este apartado se describen las técnicas de evaluación que hemos aplicado, así como los resultados cuantitativos de la misma y su discusión. Los dos aspectos más importantes de la evaluación son la construcción de la colección de prueba y la elección de las métricas de evaluación.

5.1 Colección de evaluación

Hemos construido una colección de evaluación consistente en 109 noticias obtenidas en la edición electrónica del diario ABC electrónico, para un día determinado. También hemos seleccionado tres perfiles de usuarios reales del sistema Hermes, atendiendo a que sean relativamente distintos entre sí, y a que posean distintos grados de elaboración. De este modo, el primer usuario está interesado en información deportiva y en Internet, y posee un perfil relativamente pobre. El segundo usuario posee un perfil más rico, concentrado en temas económicos y políticos, mientras que el tercer usuario posee un perfil muy completo que cubre temas económicos como la bolsa.

Un experto independiente ha examinado cada noticia, decidiendo si es relevante o no de acuerdo a los intereses mostrados en cada perfil. Se han encontrado 38, 52 y 78 noticias relevantes para cada perfil respectivamente.

5.2 Métricas de calidad

A fin de determinar la calidad de una serie de resúmenes desde el punto de vista de lo indicativos que son para juzgar el interés de las noticias enviadas por un sistema, es razonable comparar los resultados de la recuperación de las noticias completas con respecto a un perfil con los resultados de la recuperación de las mismas cuando sólo se utiliza el resumen. Este tipo de evaluación cuantitativa es característico de sistemas de extracción de resúmenes que se enmarcan dentro de un sistema más complejo que realiza otras funciones, como un sistemas de recuperación o de filtrado de información como el nuestro.

La métrica de evaluación más popular en el marco de la recuperación de información es la precisión media sobre once niveles de recall [15]. Esta métrica permite comparar numérica y gráficamente (por medio de una gráfica recall-precisión) distintos enfoques de recuperación. Nosotros calculamos la precisión media y la gráfica recall-precisión de los siguientes enfoques:

- La utilización de las noticias completas en la recuperación.
- La utilización de los resúmenes generales en la recuperación, con los parámetros $\alpha = 0.5$ y $\beta = 0.5$ (es decir, se da igual importancia a las dos heurísticas generales).
- La utilización de los resúmenes personalizados en la recuperación, con los parámetros $\alpha = 0.5$ y $\beta = 0.5$, y $\gamma = 1, 2, ..., 5$ (es decir, la heurística de personalización es igual de importante que las dos heurísticas generales, o el doble, etc.).

En general, cuanto más próxima se encuentra la precisión media o la gráfica obtenida para un

método de extracción de resúmenes a la obtenida usando todo el texto de las noticias, la calidad es mayor porque se retiene más información. La herramienta de recuperación utilizada es el motor de selección de noticias para su envío a los usuarios en el proyecto Hermes.

5.3 Resultados y discusión

En la figura 1 presentamos un grafo recall-precisión comparando los siete enfoques anteriores, calculada en media para los tres usuarios. En la tabla 1 se muestran los valores de la precisión media sobre once niveles de recall para cada uno de los enfoques comparados, sobre cada usuario y en media sobre éstos.

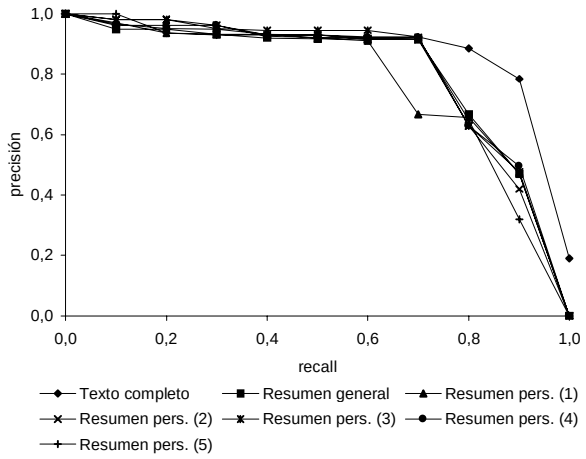

Figura 1*.Gráfica recall-precisión para los enfoques evaluados Las etiquetas de la forma "Resumen pers.(i)" identifican los resultados correspondientes a los resúmenes personalizados con valor de γ = i.*

	n	rg	rp(1)	rp(2)	rp(3)	rp(4)	rp(5)
u(1)	0,871	0,860	0,861	0,850	0,854	0,859	0,813
u(2)	0,957	0,905	0,900	0,890	0,902	0,902	0,905
u(3)	0,745	0,591	0,527	0,622	0,638	0,611	0,610
Media	0,858	0,785	0,763	0,788	0,798	0,791	0,776

Tabla 1*. Valores de precisión medios para cada uno de los enfoques comparados y para cada uno de los usuarios y en media. Las etiquetas n, rg y rp(i) identifican los resultados obtenidos utilizando el texto completo de las noticias, los resúmenes generales y los personales con valor de γ = i, respectivamente. La etiqueta u(i) identifica los resultados para el usuario iésimo.*

En la figura 1 se puede observar que para valores bajos y medios de recall, no existe diferencia perceptible entre la utilización del texto completo frente al uso de cualquier tipo de resumen en la recuperación, aunque los resultados son levemente mejores para los resúmenes personalizados con γ = 3. Para valores altos de recall, es decisivamente preferible utilizar el texto completo frente a utilizar un resumen en la recuperación, aunque los resultados son poco homogéneos.

En la tabla 1 se observa gran variación de resultados dependiendo del usuario, aunque en general, los resúmenes personalizados son levemente mejores en media para γ = 3. Sin embargo, no existe un valor de γ óptimo para todos los usuarios, dependiendo éste de la cantidad de información disponible en el perfil, y de la similitud existente entre el perfil del usuario y las noticias disponibles. Por ejemplo, aunque el tercer usuario posee gran cantidad de información en su perfil, ésta es poco útil para seleccionar noticias porque incluye varios nombres propios, que raramente aparecen en las noticias.

En general, los resultados son muy alentadores aunque las diferencias son poco significativas. Es aconsejable realizar una evaluación con un mayor número de usuarios y noticias, que muy probablemente mostrará que es preferible la utilización de resúmenes personalizados frente a resúmenes generales en sistemas de recuperación y filtrado de información.

6. *Conclusiones y trabajo futuro*

En este artículo se ha presentado un sistema de elaboración de resúmenes personalizados en el marco de un sistema de envío personalizado de noticias llamado Hermes. El sistema de extracción de resúmenes genera resúmenes adaptados a las necesidades de información del usuario de acuerdo con un perfil en Hermes. Los resúmenes obtenidos tienen como fin ayudar al usuario en la decisión de leer o no el texto completo de la noticia.

Los resultados de la evaluación de la técnica de extracción de resúmenes personalizados vs. resúmenes no personalizados son alentadores. En nuestro trabajo futuro, planeamos mejorar las técnicas de extracción de resúmenes personalizados utilizando nuevas heurísticas y

la información de recursos léxicos como WordNet y EuroWordNet. Asimismo, realizaremos una evaluación de la calidad de los resúmenes personalizados basada en técnicas orientadas al usuario y la tarea. Finalmente, portaremos el sistema al idioma inglés, tarea que plantea un coste significativamente pequeño, y evaluaremos nuestras técnicas de extracción de resúmenes en el contexto multilingüe.

Bibliografía

[1] Amato, G., Straccia, U.: User Profile Modeling and Applications to Digital Libraries. In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Paris, France (1999)

[2] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. ACM Press Books, New York.

[3] Billsus, D., Pazzani, M.J.: A Hybrid User Model for News Story Classification. In: Proceedings of the Seventh International Conference on User Modeling, Banff, Canada (1999)

[4] Buenaga, M., Gómez, J.M. and Díaz B. (1997) Using WordNet to Complement Training Information in Text Categorization. In Proceedings of the Second International Conference on Recent Advances in Natural Language Processing (RANLP).

[5] Chen, L., Sycara, K.P.: WebMate: A Personal Agent for Browsing and Searching. In: Proceedings of the Second International Conference on Autonomous Agents, Minneapolis, (1998)

[6] Díaz Esteban, A., Gervás Gómez-Navarro, P., García Jiménez, A.: Evaluating a User-Model Based Personalisation Architecture for Digital News Services. In: Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries, Lisbon, Portugal (2000)

[7] Edmundson, H.P. 1969. New Methods in Automatic Abstracting. *Journal of the ACM*, 16(2):264-285.

[8] Hahn, Udo and Inderjeet Mani. 2000. The Challenges of Automatic Summarization. *Computer*, 33(11):29-36.

[9] Hahn, Udo and Ulrich Reimer. 1999. Knowledge-Based Text Summarization: Salience and Generalization Operators for Knowledge-Based Abstraction. In Mani, I. and Maybury, M.T. (eds.), *Advances in Automatic Text Summarization*, 215-232. The MIT Press, Cambridge, Massachusetss.

[10] Kupiec, Julian, Jan O. Pedersen, and Francine Chen. 1995. A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 68-73, Seattle, Washington.

[11] Mani, Inderjeet and Mark T. Maybury (eds). 1999. *Advances in Automatic Text Summarization*. The MIT Press, Cambridge, Massachusetss.

[12] Maña, Manuel J., Manuel de Buenaga, and José María Gómez. 1999. Using and Evaluating User Directed Summaries to Improve Information Access. In S. Abiteboul and A.M. Vercoustre (eds.), *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)*, Lecture Notes in Computer Science, Vol. 1696, 198-214, Springer-Verlag.

[13] Porter, M.F. An algorithm for suffix stripping. Program, 14(3): 130-137

[14] Salton, G., J. Allan, C. Buckley, and A. Singhal. 1994. Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts. *Science*, 264, 1421-1426.

[15] Salton, G. (1989). Automatic Text Processing: the transformation, analysis and retrieval of information by computer. Addison Wesley. 1989

[16] Sebastiani, F. "A Tutorial on Automated Text Categorisation". Proceedings of the First Argentinean Symposium on Artificial Intelligence (ASAI-99)

[17] Teufel, Simone and Marc Moens. 1997. Sentence extraction as a classification task. En *Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization*, Madrid.

[18] Yan, T.W., Garcia-Molina, H.: SIFT – A Tool for Wide-Area Information Dissemination. In: Proceedings of the USENIX Technical Conference, (1995)